Céspedes-Pablo, Javier; Olmos, Pablo M.; Sánchez-Fernández, Matilde; Pérez-Cruz, Fernando. (2018). Probabilistic MIMO symbol detection with expectation consistency approximate inference. *IEEE Transactions on Vehicular Technology*, 67(4), pp. 3481–3494.

DOI: 10.1109/TVT.2017.2786638

Probabilistic MIMO Symbol Detection with Expectation Consistency Approximate Inference

Javier Céspedes, Pablo M. Olmos, Matilde Sánchez-Fernández, Fernando Perez-Cruz

***Abstract*—In this paper we explore low-complexity probabilistic algorithms for soft symbol detection in high-dimensional multiple-input multiple-output (MIMO) systems. We present a novel algorithm based on the Expectation Consistency (EC) framework, which describes the approximate inference problem as an optimization over a non-convex function. EC generalizes algorithms such as Belief Propagation and Expectation Propagation. For the MIMO symbol detection problem, we discuss feasible methods to find stationary points of the EC function and explore their tradeoffs between accuracy and speed of convergence. The accuracy is studied, first in terms of input-output mutual information and show that the proposed EC MIMO detector greatly improves state-of-the-art methods, with a complexity order cubic in the number of transmitting antennas. Second, these gains are corroborated by combining the probabilistic output of the EC detector with a low-density parity-check (LDPC) channel code.**

***Index Terms*—MIMO Communication Systems, Approximate Inference, Expectation Consistency, Low-density Parity-Check Codes.**

I. Introduction

With the increasing demand for higher data rates, multiple-input multiple-output (MIMO) systems have attracted much attention over the last decade [1]. It is well known that MIMO communication systems achieve substantial gains in terms of spectral efficiency compared to conventional single-input single-output (SISO) systems. In fact, it has been shown that under ideal conditions the capacity of a point-to-point MIMO system with m transmitting antennas and r receiving antennas scales linearly with $\min(m, r)$, which is referred to as the multiplexing gain [2].

Modern channel-coding techniques, such as Turbo codes [3] or LDPC codes [4], are needed to achieve transmission rates close to the fundamental theoretical limits of the MIMO channel. Efficient decoding is possible using the belief propagation (BP) algorithm [4], [5], which is a low-complexity message-passing approximate inference method to estimate marginals in a joint probability distribution. BP decoding needs as input an estimate to the posterior probability of each coded bit given the vector of channel observations. This information is provided by the so-called probabilistic symbol detector, which has to marginalize the posterior probability density function (pdf) of the transmitted vector of symbols, given the channel observation. For a MIMO channel, this has complexity $\mathcal{O}(M^m)$, where M is the constellation order.

Multiple algorithms have been proposed to perform hard-output symbol detection in MIMO systems, see [6]–[15]. On the contrary, the list of probabilistic symbol detection algorithms is comparatively much shorter. Soft-ouput sphere decoding (SD) methods solve the marginalization in a subspace of the constellation alphabet $\mathcal{A}^m$ [16]–[18]. However, to maintain good performance, the dimension of the subspace must grow rapidly with m, the modulation order and the inverse of the signal-to-noise ratio (SNR) [19]. Thus, SD methods are not suitable for massive MIMO scenarios, where both m and M are potentially very large. Alternatively, some other works consider the use of Markov chain Monte Carlo (MCMC) algorithms to approximate the marginal posterior probabilities [20]–[23]. While this approach has been shown to be viable for hard-output symbol detection, probabilistic detection requires a sufficiently large number of samples *per constellation point* at each transmitter. For large m and high-order constellations, MCMC methods become excessively burdensome.

The focus of this paper is on MIMO probabilistic symbol detection methods that can scale up to hundreds of antennas and high-order modulations based on quadrature amplitude modulation (QAM). In particular, we focus on methods with polynomial complexity with the number m of transmit antennas. The minimum-mean-squared error (MMSE) solution can be cast as a probabilistic detector since it computes the mode of a Gaussian approximation to the posterior pdf of the MIMO symbols [13], [24], likewise its soft successive interference cancellation (soft MMSE-SIC) version [25]. In both implementations complexity is dominated by an $m \times m$ matrix inversion. The Gaussian tree approximation (GTA) algorithm [26], very close in hard detection performance to MMSE-SIC, is a detection algorithm that constructs a tree-factorized approximation to posterior pdf of the symbols, to then estimate marginals distributions using BP. Also, inspired by their success in compressed sensing [27], in recent years there has been an intense research interest on MIMO detection techniques based on message passing algorithms. We can mention the channel hardening-exploiting message passing (CHEMP) in [28] and the Gaussian Message Passing Iterative Detector (GMPID) in [29], [30]. Both methods have been shown to be effective (close to SD methods) for large

This work has been partly funded by the Spanish Government through projects MIMOTEX (TEC2014-61776-EXP), CIES (RTC-2015-4213-7), ELISA (TEC2014-59255-C3-3R) and FLUID (TEC2016-78434-C3-3-R), by the Juan de la Cierva program (IJCI-2014-19150) and by Comunidad de Madrid (project 'CASI-CAM-CM', id. S2013/ICE-2845)

J. Céspedes, P. M. Olmos and M. Sánchez-Fernández are with the Signal Theory & Communications Department, Universidad Carlos III de Madrid. Pablo M. Olmos is also with the Gregorio Marañón Health Research Institute. E-mail: {jcespedes,olmos,mati}@tsc.uc3m.es

F. Perez-Cruz is with the Signal Theory & Communications Department, Universidad Carlos III de Madrid and is an Associate Professor at Stevens Institute of Technology. E-mail: fernando@tsc.uc3m.es

MIMO systems with QPSK constellations. However, asymptotic analysis of this type of algorithms shows that they do not perform well with high-order QAM constellations unless the number r of receiving antennas is much larger than the number m of transmitting antennas [30], [31]. An improved version of the GMPID algorithm called SA-GMPID is shown to asymptotically converge to the MMSE detection solution even for the case $m/r > 1$ [32]. We remark that in this paper we propose algorithms that, while having larger complexity compared to these type of message-passing algorithms, they significantly improve the MMSE solution.

In [33], we proposed Expectation Propagation (EP) [34], [35] to perform hard-output MIMO symbol detection in the high SNR regime. In that paper, EP is used to find the mode of the posterior probability distribution by projecting it into a Gaussian approximation. The method cannot be easily generalized to perform probabilistic detection, as its description is essentially an iterative algorithm that does not provide the complete picture of the fundamental underlying inference problem. Actually, in [36] we showed that, while the MIMO EP receiver in [33] is able to significantly improve GTA as hard detector, achieving gains of around 2 dBs, both methods perform similarly when combined with an LDPC channel decoder that requires a probabilistic input. In a simpler scenario, i.e. channel equalization for single-user intersymbol-interference (ISI) channels, different heuristics have been recently proposed in [37] to improve the EP probabilistic output, but it is shown that ultimately a turbo-like receiver, where the LDPC decoder output is fed back to the EP equalizer, is required to obtain a robust solution that is not tailored to a particular modulation or channel instance.

In this work, we consider one-shot receiver architectures, in which the channel decoder output is not fed back to the MIMO symbol detector to modify the original estimate. In this scenario, the design of the MIMO detector is particularly crucial, as the overall system performance highly depends on its accuracy. One-shot receivers can be used in latency-constrained applications instead of iterative Turbo-like receivers, as the latency in the latter case can become too large if long block channel codes are used [38]. Furthermore, we show how probabilistic MIMO symbol detection can be implemented using a general approximate inference framework called Expectation Consistency (EC), which was first described by Opper & Winther in [39]. In EC, we describe the inference problem as the search of a stationary point of an approximation to the free energy associated to the true posterior probability distribution of the transmitted symbols. Any stationary point satisfies a moment matching condition between the involved distributions. In this paper, we tailor the original EC formulation to the MIMO detection case and we discuss feasible methods to find such stationary points and show the fundamental tradeoffs between accuracy and speed of convergence. In particular, we propose an update rule that performs very close to the moment matching EC solution, with a complexity comparable to running MMSE ten times. Also, we propose methods to overcome numerical instabilities that may arise in the MIMO detection scenario, particularly when we use large constellation alphabets. In all tested scenarios, we find solutions that are robust and accurate across different modulation orders and system dimensions. Finally, the resulting EC probabilistic MIMO detector achieves excellent performance results compared to state-of-the-art methods with the same complexity order.

To measure the accuracy of the EC MIMO detector probabilistic output, first we use a Monte Carlo estimate to the mutual information between the transmitted MIMO symbol vector and the corresponding output of the probabilistic symbol detection stage. At high SNRs, all detection methods saturate at the same mutual information level, i.e., $\log_2(M)$ bits per channel use per antenna, due to the use of a finite discrete constellation of M points. Operating in the high-SNR region of saturation is undesirable, as the gap to channel capacity grows exponentially as we increase the SNR. However, at moderate SNR, our proposed detector outperforms other detectors in the literature and, in those scenarios where we could obtain the optimal detector solution, EC gets very close to it. Second, the predicted gain at moderate-SNRs is corroborated by bit error rate (BER) performance simulation using optimized irregular LDPC block codes [40] and terminated convolutional-LDPC block codes [41], [42]. In all cases, we obtain remarkable SNR gains, proving that the accuracy of the MIMO probabilistic symbol detection stage is crucial in the system's performance.

Overall, the contributions of this paper are summarized as follows:

- We introduce EC approximate inference framework and show how it can be applied to the MIMO detection scenario, developing the EC free energy approximation and computing its gradients.
- We compare several approaches to find EC stationary points, and propose iterative rules that are able to approach the optimal solution at $\mathcal{O}(m^3)$ complexity.
- We obtain the achievable rate (mutual information) of a single-user MIMO system to show the accuracy in the pdf approximation to the true posterior, also proving that with EC detection we significantly reduce the gap to capacity. The predicted gains are corroborated via error rate simulation with optimized LDPC codes.

The paper is structured as follows. In Section II we review the system model. In Section III, we discuss on the transmission rate and how it depends on the MIMO symbol detection method implemented, highlighting the importance of a good approximation to the true posterior. Section IV briefly presents the EC approximate inference framework and we tailor it to the MIMO detection case in Section V. In Section VI, experimental results are presented. Final conclusions and potential lines of future research are described in Section VII.

Notation: Capital and lowercase boldface symbols represent matrices and vectors respectively. $[\cdot]^\top$ is the transpose and $[\cdot]^H$ is the Hermitian. Finally, $[n]$ denotes the set $\{1, 2, \ldots, n\}$.

II. System Model

Consider a single-user MIMO system where m transmitting antennas communicate to a receiver with r antennas. The system model is shown in Fig. 1. Let $\mathbf{b} = [b_1, b_2, ..., b_k]^\top$ denote the input information binary vector, which is Gray-mapped

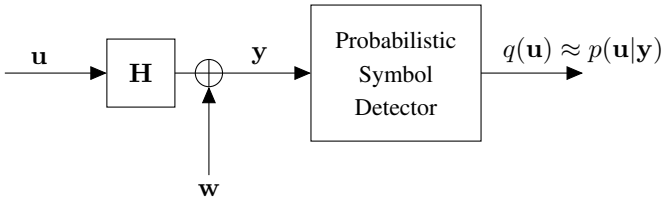


Fig. 1. System model

and modulated into QAM symbols. Then, an m-dimensional vector of QAM symbols is generated, that is denoted by $\mathbf{u} = \mathbf{u}_{\text{re}} + j\mathbf{u}_{\text{im}} \in \mathcal{A}^m$, where $|\mathcal{A}| = M$. The symbol vector $\mathbf{u}$, is transmitted over a memoryless flat-fading complex MIMO channel, defined as a matrix $\mathbf{H}$ with dimensions $r \times m$ of zero-mean unit-variance complex Gaussian coefficients. Therefore,

$$\mathbf{y} = \mathbf{H}\mathbf{u} + \mathbf{w}, \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^r$ and $\mathbf{w} \in \mathbb{C}^r$ is an additive white circular-symmetric complex Gaussian noise vector with independent zero-mean components and σ_w^2-variance. We also assume that the receiver has perfect channel state information (CSI). On the other hand, the signal-to-noise ratio is defined as

$$\text{SNR(dB)} = 10\log_{10}\left(m\log_2(M)\frac{E_b}{\sigma_w^2}\right), \tag{2}$$

where E_b is the bit energy and the constellation energy E_s can be written as

$$E_s = E_b\log_2(M). \tag{3}$$

Note that the SNR defined is taking into account the full power transmission instead of the per-antenna power. Given the channel observation, the posterior distribution of the transmitted symbols, that would lead to the optimal detector and that is also denoted through this work as true posterior, is

$$p(\mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{u})p(\mathbf{u})}{p(\mathbf{y})} \propto \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I})p(\mathbf{u}), \tag{4}$$

where $\mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I})$ denotes a complex Gaussian with mean $\mathbf{H}\mathbf{u}$ and covariance matrix $\sigma_w^2\mathbf{I}$, and $p(\mathbf{u})$ is the prior probability density function for $\mathbf{u}$. Assuming that we transmit independent uniformly distributed symbols, we have

$$p(\mathbf{u}) = \prod_{i=1}^{m} p(u_i) = \prod_{i=1}^{m} \frac{1}{M}\mathbb{I}_{u_i \in \mathcal{A}}, \tag{5}$$

where $\mathbb{I}_{u_i \in \mathcal{A}}$ takes value one if u_i belongs to $\mathcal{A}$. Observe that, due to the likelihood term in (4), $p(\mathbf{u}|\mathbf{y})$ is a multidimensional discrete distribution that maps over a fully connected factor graph. Exact inference over $p(\mathbf{u}|\mathbf{y})$, required to evaluate symbol marginals $p(u_i|\mathbf{y})$, $i \in [m]$, to later feed a modern channel decoder, has cost $\mathcal{O}(M^m)$ and quickly (in both M and m) becomes unfeasible.

A. Posterior approximation and inference

One of the alternatives to implement a low complexity probabilistic symbol detector is to construct a tractable distribution $q(\mathbf{u})$ that approximates $p(\mathbf{u}|\mathbf{y})$. By tractable we mean that performing inference over $q(\mathbf{u})$, namely marginalizing it or computing expectations, is feasible. Other options, reduce or modify the constellation space, as for example SD.

Focusing on the first alternative, the MMSE method can be seen as a Gaussian approximation $q(\mathbf{u})$ to $p(\mathbf{u}|\mathbf{y})$ obtained by replacing the independent discrete priors in (5) by the product of univariate zero-mean and E_s-variance complex circularly-symmetric Gaussian factors [13], [24]. The Gaussian tree approximation (GTA) was first proposed in [26]. The method constructs a tractable cycle-free discrete approximation to (4) by replacing the Gaussian likelihood term $p(\mathbf{y}|\mathbf{u})$ by a Gaussian distribution that factorizes in cycle-free graph, chosen to match the marginal and cross-moments of $p(\mathbf{y}|\mathbf{u})$. Using this cycle-free approximation to the likelihood, efficient inference is carried out using BP. Finally, there exist several recent proposals that perform approximate inference for MIMO symbol detection based on approximate message passing (AMP) [27]. AMP algorithms essentially implement the standard rules of BP message passing [43] and all messages are approximated with univariate Gaussian distributions. Among AMP methods for MIMO detection, we can mention the CHEMP algorithm in [28] and GMPID in [30]. An approximation to $p(\mathbf{u}|\mathbf{y})$ can be constructed from the AMP marginals using the Bethe reparameterization [43].

In Section V-D, we have included a table summarizing the theoretical complexity order of each of the MIMO detection methods we use in our experiments.

III. Transmission rate

Consider a fixed and known channel matrix $\mathbf{H}$, under the system model defined in Section II. With the power constraint $\mathbb{E}[\mathbf{u}^T\mathbf{u}] \leq \text{SNR}\sigma_w^2$, the ergodic channel capacity per transmitted antenna with perfect CSI at the receiver and no CSI at the transmitter is given by

$$C = \max_{p(\mathbf{u})} \frac{I(\mathbf{u},\mathbf{y})}{m} = \frac{\log_2(\det(\mathbf{I}_r + \frac{\text{SNR}}{m}\mathbf{H}\mathbf{H}^H))}{m} \tag{6}$$

bits per channel use and antenna. Capacity is achieved when $\mathbf{u}$ is Gaussian distributed with zero-mean and covariance matrix equal to identity [44].

When $\mathbf{u}$ is a random vector uniformly distributed in $\mathcal{A}^m$, the system transmission rate degrades and can be far from the capacity limit in (6). The achievable rate per antenna can be computed by evaluating the mutual information between u_i, the transmitted symbol at i-th antenna and $\hat{u}_i \sim p(u_i|\mathbf{y})$, i.e.,

$$I(u_i,\hat{u}_i) = \mathbb{E}_{p(u_i,\hat{u}_i)}\left[\log_2\frac{p(\hat{u}_i|u_i)}{p(\hat{u}_i)}\right] \quad \text{(bits/channel use)}, \tag{7}$$

for $i \in [m]$. Unfortunately, it is not possible to compute this mutual information in closed-form. We follow a Monte Carlo procedure to estimate $\frac{1}{m}\sum_{i=1}^{m} I(u_i,\hat{u}_i)$ in the same channel knowledge scenario as the one assumed in (6), namely perfect CSI only at the receiver. More precisely, we estimate $I(u_i,\hat{u}_i)$,

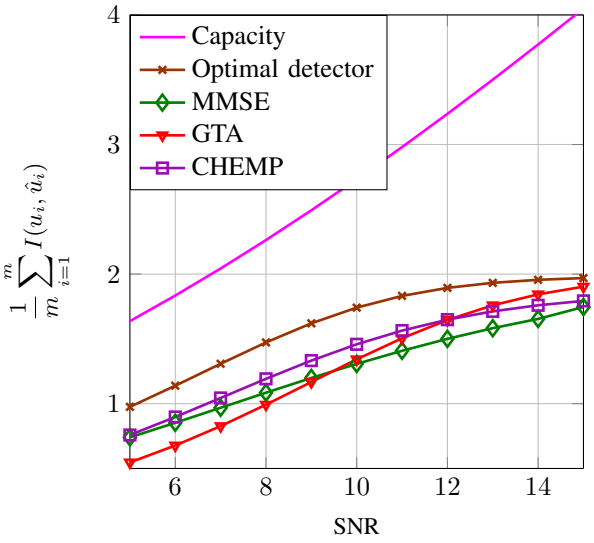


Fig. 2. Transmission rate in 5×5 scenario with QPSK modulation.

$i \in [m]$, at one particular SNR point as follows: first, we collect $N \in \mathbb{Z}_+$ samples from the joint distribution of $u_i, \mathbf{y}$ and $\hat{u}_i$. Using this set of samples, we estimate $p(\hat{u}_i)$, $p(\hat{u}_i|u_i)$ for any $u_i, \hat{u}_i \in \mathcal{A}$, and, finally, compute a numerical estimate to $I(u_i, \hat{u}_i)$ in (7). As $N \to \infty$, the estimate to $I(u_i, \hat{u}_i)$ gets tight. Samples of the joint $(\mathbf{u}, \mathbf{y}, \hat{\mathbf{u}})$ distribution are computed using *ancestral sampling* [45]. Each of the N samples is generated following the next steps:

1) Sample $\mathbf{u}$ from an uniform distribution in $\mathcal{A}^m$.
2) Sample $\mathbf{y}$ from $p(\mathbf{y}|\mathbf{u}, \mathbf{H})$.
3) Sample $\hat{u}_i$, $i \in [m]$, from

$$p(u_i|\mathbf{y}) = \sum_{\mathbf{u}_{-i}} p(\mathbf{u}|\mathbf{y}) \qquad u_i \in \mathcal{A}, \tag{8}$$

where $\mathbf{u}_{-i}$ denotes all elements in $\mathbf{u}$ except u_i.

When a probabilistic symbol detector does not use the true posterior, the transmission rate can be evaluated by following a similar procedure, but in 3) we sample $\hat{u}_i$ after marginalization over $q(\mathbf{u})$, namely the approximation constructed to $p(\mathbf{u}|\mathbf{y})$. Thus, the average mutual information computed for each low complexity detection method is used as a performance metric that measures how close $q(\mathbf{u})$ is to $p(\mathbf{u}|\mathbf{y})$. At the same time, the better the quality of the approximation is, the higher the rate becomes. Note also that to compute this metric, we consider uncoded transmission. For instance, Fig. 2 shows the average mutual information per antenna in a 5×5 scenario with QPSK modulation for both the optimal detector (which works directly with the true posterior $p(\mathbf{u}|\mathbf{y})$), and for MMSE, GTA and CHEMP suboptimal detectors. It has been computed with $N = 10^6$ samples per SNR point. Also, results have been averaged over 100 realizations of $\mathbf{H}$. Observe that all methods operate close to the limit of 2 bits/channel use when the SNR is high, but the gap to channel capacity in this regime grows exponentially fast with the SNR. For intermediate SNR values, optimal detection clearly outperforms MMSE, GTA and CHEMP detection[1] . It is precisely in this regime where we must improve the accuracy of the probabilistic symbol detection stage.

IV. Expectation Consistency Approximate Inference for MIMO detection

In this section we give a brief introduction to EC approximate inference [39], to then tailor it for low-complexity probabilistic MIMO detection. Let $\boldsymbol{U}$ be a random variable with a probability density function that factors in the following way

$$p(\mathbf{u}) = \frac{1}{Z} f_q(\mathbf{u}) f_r(\mathbf{u}), \tag{9}$$

where we assume that computing $Z = \int f_q(\mathbf{u}) f_r(\mathbf{u}) d\mathbf{u}$ or any expectation w.r.t. $p(\mathbf{u})$ is unfeasible. However, we do assume that, separately, $f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ are tractable w.r.t. a measure of the form $\exp(\boldsymbol{\lambda}^T \boldsymbol{\phi}(\mathbf{u}))$ for some function vector $\boldsymbol{\phi}(\mathbf{u}) = [\phi_1(\mathbf{u}), \ldots, \phi_J(\mathbf{u})]$. Namely, we assume it is possible to perform inference over the following two distributions used to approximate $p(\mathbf{u})$:

$$q(\mathbf{u}) = \frac{1}{Z_q(\boldsymbol{\lambda}_q)} f_q(\mathbf{u}) \exp(\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})), \tag{10}$$

$$r(\mathbf{u}) = \frac{1}{Z_r(\boldsymbol{\lambda}_r)} f_r(\mathbf{u}) \exp(\boldsymbol{\lambda}_r^\top \boldsymbol{\phi}(\mathbf{u})), \tag{11}$$

where the $J \times 1$ parameter vectors $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ belong to a certain convex set Φ, and

$$Z_q(\boldsymbol{\lambda}_q) = \int f_q(\mathbf{u}) \exp(\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u}, \tag{12}$$

$$Z_r(\boldsymbol{\lambda}_r) = \int f_r(\mathbf{u}) \exp(\boldsymbol{\lambda}_r^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u}. \tag{13}$$

Note that both $q(\mathbf{u})$ and $r(\mathbf{u})$ define an exponential family of distributions[2], where $\boldsymbol{\lambda}_q$ ($\boldsymbol{\lambda}_r$) is the natural parameter vector, $\boldsymbol{\phi}(\mathbf{u})$ is the vector of sufficient statistics and $\log Z_q(\boldsymbol{\lambda}_q)$ ($\log Z_r(\boldsymbol{\lambda}_r)$) is a convex function of $\boldsymbol{\lambda}_q$ ($\boldsymbol{\lambda}_r$) that satisfies

$$\nabla_{\boldsymbol{\lambda}_q} \log Z_q(\boldsymbol{\lambda}_q) = \mathbb{E}_{q(\mathbf{u})} \left[\boldsymbol{\phi}(\mathbf{u})\right], \tag{14}$$

$$\nabla_{\boldsymbol{\lambda}_r} \log Z_r(\boldsymbol{\lambda}_r) = \mathbb{E}_{r(\mathbf{u})} \left[\boldsymbol{\phi}(\mathbf{u})\right]. \tag{15}$$

The main idea behind EC approximate inference is to optimize $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_r$ so that $q(\mathbf{u})$ and $r(\mathbf{u})$ have the same moments, i.e., (14) is consistent with (15), keeping in mind that both $q(\mathbf{u})$ and $r(\mathbf{u})$, being the functions used to approximate $p(\mathbf{u})$, contain "partial information" ($f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ respectively) of this true distribution $p(\mathbf{u})$.

The first step to derive the EC approximation is to note that the partition function Z in (9) can be expressed the following way

$$Z = Z_q(\boldsymbol{\lambda}_q) \frac{Z}{Z_q(\boldsymbol{\lambda}_q)} = Z_q(\boldsymbol{\lambda}_q) \frac{\int f_q(\mathbf{u}) f_r(\mathbf{u}) d\mathbf{u}}{Z_q(\boldsymbol{\lambda}_q)} =$$

$$= Z_q(\boldsymbol{\lambda}_q) \int \frac{f_q(\mathbf{u})}{Z_q(\boldsymbol{\lambda}_q)} f_r(\mathbf{u}) \exp((\boldsymbol{\lambda}_q - \boldsymbol{\lambda}_q)^\top \boldsymbol{\phi}(\mathbf{u})) d\mathbf{u} \tag{16}$$

$$= Z_q(\boldsymbol{\lambda}_q) \mathbb{E}_{q(\mathbf{u})}[f_r(\mathbf{u}) \exp(-\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u}))]. \tag{17}$$

[1]Note the similarities with the throughput results presented in [46].
[2]See [43] for an introduction to exponential families and their properties.

And thus,

$$\log Z = \log Z_q(\boldsymbol{\lambda}_q) + \log\left(\mathbb{E}_{q(\mathbf{u})}[f_r(\mathbf{u})\exp(-\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u}))]\right). \tag{18}$$

where $\log Z$ is also known as the *energy function*. In order to estimate the expectation in the above expression, we replace $q(\mathbf{u})$ by a simpler distribution $s(\mathbf{u})$ that belongs to the same exponential family than $q(\mathbf{u})$ and $r(\mathbf{u})$, i.e.,

$$s(\mathbf{u}) = \frac{1}{Z_s(\boldsymbol{\lambda}_s)}\exp(\boldsymbol{\lambda}_s^\top \boldsymbol{\phi}(\mathbf{u})), \tag{19}$$

where $\log Z_s(\boldsymbol{\lambda}_s)$ is a convex function of $\boldsymbol{\lambda}_s$ that satisfies $\nabla_{\boldsymbol{\lambda}_s}\log Z_s(\boldsymbol{\lambda}_s) = \mathbb{E}_{s(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})]$. While replacing $q(\mathbf{u})$ by $s(\mathbf{u})$ yields, in general, a poor approximation, it can be a fairly reasonable solution if both $q(\mathbf{u})$ and $s(\mathbf{u})$ have the same moments, namely if $\mathbb{E}_{q(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})] = \mathbb{E}_{s(\mathbf{u})}[\boldsymbol{\phi}(\mathbf{u})]$. This condition is naturally achieved as a stationary point of the resulting approximation to $\log Z$. By replacing $q(\mathbf{u})$ by $s(\mathbf{u})$ in (18), $\log Z$ is approximated by

$$\begin{aligned}&\log Z_{\text{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) = \\ &= \log Z_q(\boldsymbol{\lambda}_q) + \log(\mathbb{E}_{s(\mathbf{u})}[f_r(\mathbf{u})\exp(-\boldsymbol{\lambda}_q^\top \boldsymbol{\phi}(\mathbf{u}))]),\end{aligned} \tag{20}$$

and after simple manipulation this term can be expressed as follows:

$$\begin{aligned}&\log Z_{\text{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s) = \\ &= \log Z_q(\boldsymbol{\lambda}_q) + \log Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q) - \log Z_s(\boldsymbol{\lambda}_s).\end{aligned} \tag{21}$$

Recall that by assumption $Z_q(\boldsymbol{\lambda}_q)$, $Z_r(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)$ and $Z_s(\boldsymbol{\lambda}_s)$ can be computed efficiently. And note that $\log Z_{\text{EC}}$ depends only on $\boldsymbol{\lambda}_q$ and $\boldsymbol{\lambda}_s$, while it depends on three probability distributions: $q(\mathbf{u})$ with parameter vector $\boldsymbol{\lambda}_q$, $r(\mathbf{u})$ with parameter vector $(\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_q)$ and $s(\mathbf{u})$ with parameter vector $\boldsymbol{\lambda}_s$. Recall we seek moment matching between $q(\mathbf{u})$ and $r(\mathbf{u})$ and also between $q(\mathbf{u})$ and $s(\mathbf{u})$. While the first condition ensures that the two approximations that we construct to $p(\mathbf{u})$ are consistent, the latter is required so that the measure replacement in the expectation in (18) is not too coarse. Both conditions are satisfied at any point $(\boldsymbol{\lambda}_q^*, \boldsymbol{\lambda}_s^*)$ where the gradient of the EC energy function $\log Z_{\text{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$ is zero, i.e. optimization over $\log Z_{\text{EC}}(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_s)$ would lead to $(\boldsymbol{\lambda}_q^*, \boldsymbol{\lambda}_s^*)$.

A. The EC free energy for MIMO detection

To simplify the low-complexity detector derivation, we rewrite the probabilistic model in (4) to work with real-valued distributions, considering the real $\mathcal{R}(\cdot)$ and imaginary $\mathcal{I}(\cdot)$ parts separately. Define $\widetilde{\mathbf{u}} = \left[\mathbf{u}_{\text{re}}^\top \quad \mathbf{u}_{\text{im}}^\top\right]^\top$, $\widetilde{\mathbf{y}} = \left[\mathcal{R}(\mathbf{y})^\top \quad \mathcal{I}(\mathbf{y})^\top\right]^\top$, $\widetilde{\mathbf{w}} = \left[\mathcal{R}(\mathbf{w})^\top \quad \mathcal{I}(\mathbf{w})^\top\right]^\top$ and

$$\widetilde{\mathbf{H}} = \begin{bmatrix}\mathcal{R}(\mathbf{H}) & -\mathcal{I}(\mathbf{H}) \\ \mathcal{I}(\mathbf{H}) & \mathcal{R}(\mathbf{H})\end{bmatrix}.$$

Thus, the real-valued channel model is

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{H}}\widetilde{\mathbf{u}} + \widetilde{\mathbf{w}}, \tag{22}$$

where $\sigma_{\widetilde{w}}^2 = \sigma_w^2/2$ is the variance of the real and imaginary parts of the noise and we define $\widetilde{\mathcal{A}}$ as the new alphabet for the real and imaginary components of the M-QAM constellation, $\widetilde{\mathbf{u}} \in \widetilde{\mathcal{A}}^{2m}$, with energy $\widetilde{E}_s = E_s/2$. In the rest of this work we adopt the real-valued channel model formulation in (22) and we drop the model indicator $\widetilde{(\cdot)}$ to keep the notation uncluttered. Therefore, the a posteriori probability pdf of the transmitted symbol vector $\mathbf{u}$, and that we propose to approximate with tractable pdfs, can be expressed as follows

$$p(\mathbf{u}|\mathbf{y}) = \frac{1}{Z}\mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I})\prod_{i=1}^{2m}\mathbb{I}_{u_i\in\mathcal{A}}, \tag{23}$$

The matching of (23) with functions $f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ in (9) will be done so that $q(\mathbf{u})$ and $r(\mathbf{u})$ in (10) and (11) are tractable w.r.t. a measure of the form $\exp(\boldsymbol{\lambda}^T\boldsymbol{\phi}(\mathbf{u}))$, which means that we have to be able to easily compute moments of the form $\mathbb{E}[\boldsymbol{\phi}(\mathbf{u})]$ w.r.t. both distributions. For an EC based low-complexity detector we choose the vector of statistics and natural parameters as follows

$$\boldsymbol{\phi}(\mathbf{u}) = \left[u_1, u_2, \ldots, u_{2m}, \frac{-u_1^2}{2}, \frac{-u_2^2}{2}, \ldots, \frac{-u_{2m}^2}{2}\right]^\top, \tag{24}$$

$$\boldsymbol{\lambda} = [\gamma_1, \gamma_2, \ldots, \gamma_{2m}, \Lambda_1, \Lambda_2, \ldots, \Lambda_{2m}]^\top = [\boldsymbol{\gamma}, \boldsymbol{\Lambda}]^\top, \tag{25}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{2m}$ and $\boldsymbol{\Lambda} \in \mathbb{R}_+^{2m}$. According to (24), this choice of $\boldsymbol{\phi}(\mathbf{u})$ implies that at any zero-gradient point of the EC energy function in (21), the distributions $q(\mathbf{u})$ and $r(\mathbf{u})$ must be consistent only in their marginal first and second order moments. Under this assumption, if we choose functions $f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ as follows

$$f_q(\mathbf{u}) = \mathcal{N}(\mathbf{y} : \mathbf{H}\mathbf{u}, \sigma_w^2\mathbf{I}), \text{ and } f_r(\mathbf{u}) = \prod_{i=1}^{2m} p(u_i) \tag{26}$$

then we conclude that $q(\mathbf{u})$ and $r(\mathbf{u})$ are tractable probability density functions, since $q(\mathbf{u})$ is a Multivariate Normal distribution and $r(\mathbf{u})$ is a discrete independent distribution. More precisely, according to (10) and (26), we have

$$\begin{aligned}q(\mathbf{u}) &= \frac{1}{Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)} f_q(\mathbf{u})\exp\left(\boldsymbol{\gamma}_q^\top\mathbf{u} - \frac{\mathbf{u}^\top\operatorname{diag}(\boldsymbol{\Lambda}_q)\mathbf{u}}{2}\right) \\ &= \frac{\exp\left(\underbrace{\left(\frac{\mathbf{H}^\top\mathbf{y}}{\sigma_w^2} + \boldsymbol{\gamma}_q\right)^\top}_{\mathbf{g}^\top}\mathbf{u} - \frac{1}{2}\mathbf{u}^\top\underbrace{\left(\frac{\mathbf{H}^\top\mathbf{H}}{\sigma_w^2} + \operatorname{diag}(\boldsymbol{\Lambda}_q)\right)}_{\mathbf{S}}\mathbf{u}\right)}{Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)},\end{aligned} \tag{27}$$

where $\operatorname{diag}(\boldsymbol{\Lambda}_q)$ is a diagonal matrix with main diagonal given by $\boldsymbol{\Lambda}_q$. Therefore $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} : \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = \mathbf{S}^{-1}$ and $\boldsymbol{\mu} = \mathbf{S}^{-1}\mathbf{g}$. Also, we obtain

$$\log Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q) = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^T + \frac{1}{2}\log|\boldsymbol{\Sigma}|. \tag{28}$$

By applying standard rules for matrix derivatives, we can check that

$$\frac{\partial\log Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)}{\partial\gamma_{q,i}} = \mathbb{E}_q[u_i] = \mu_i, \tag{29}$$

$$\frac{\partial\log Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)}{\partial\Lambda_{q,i}} = -\frac{1}{2}\mathbb{E}_q[u_i^2] = -\frac{1}{2}\left(\Sigma_{ii} + \mu_i^2\right). \tag{30}$$

On the other hand, from the definition of $f_r(\mathbf{u})$ in (26) we get

$$r(\mathbf{u}) = \frac{1}{Z_r(\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r)} \exp\left(\boldsymbol{\gamma}_r^T \mathbf{u} - \frac{\mathbf{u}^T \operatorname{diag}(\boldsymbol{\Lambda}_r)\, \mathbf{u}}{2}\right) \prod_{i=1}^{2m} \mathbb{I}_{u_i \in \mathcal{A}}$$
$$= \frac{1}{Z_r(\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r)} \prod_{i=1}^{2m} \exp\left(\gamma_{ri} u_i - \frac{\Lambda_{ri} u_i^2}{2}\right) \mathbb{I}_{u_i \in \mathcal{A}}. \quad (31)$$

Therefore, $r(\mathbf{u})$ is an independent discrete pmf over $\mathcal{A}^{2m}$ such that, for $i \in [2m]$,

$$\mathbb{E}_r[u_i] = \frac{\sum_{u_i \in \mathcal{A}} u_i \exp\left(\gamma_{ri} u_i - \frac{\Lambda_{ri} u_i^2}{2}\right)}{\sum_{a \in \mathcal{A}} \exp\left(\gamma_{ri} a - \frac{\Lambda_{ri} a^2}{2}\right)}, \quad (32)$$

$$\mathbb{E}_r[u_i^2] = \frac{\sum_{u_i \in \mathcal{A}} u_i^2 \exp\left(\gamma_{ri} u_i - \frac{\Lambda_{ri} u_i^2}{2}\right)}{\sum_{a \in \mathcal{A}} \exp\left(\gamma_{ri} a - \frac{\Lambda_{ri} a^2}{2}\right)}. \quad (33)$$

Also we have

$$\log Z_r(\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r) = \log\left(\sum_{a \in \mathcal{A}} \exp\left(\gamma_{ri} a - \frac{\Lambda_{ri} a^2}{2}\right)\right), \quad (34)$$

where we can again check that, $\frac{\partial \log Z_r(\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r)}{\partial \gamma_{r,i}} = \mathbb{E}_r[u_i]$ and $\frac{\partial \log Z_r(\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r)}{\partial \Lambda_{r,i}} = -\frac{1}{2}\mathbb{E}_r[u_i^2]$, for $i \in [2m]$. Finally, the averaging distribution $s(\mathbf{u})$ in (19) is given by

$$s(\mathbf{u}) = \frac{1}{Z_s(\boldsymbol{\lambda}_s)} \exp\left(\boldsymbol{\gamma}_s^\top \mathbf{u} - \frac{\mathbf{u}^\top \operatorname{diag}(\boldsymbol{\Lambda}_s)\, \mathbf{u}}{2}\right), \quad (35)$$

and therefore $s(\mathbf{u})$ is an independent Gaussian distribution, i.e. $s(\mathbf{u}) = \mathcal{N}(\mathbf{u} : \operatorname{diag}\left(\boldsymbol{\Lambda}_s^{-1}\right) \boldsymbol{\gamma}_s, \operatorname{diag}\left(\boldsymbol{\Lambda}_s^{-1}\right))$.

Note that, given the vector of moments in (24), any choice for the functions $f_q(\mathbf{u})$ and $f_r(\mathbf{u})$ different to (26), where some discrete priors are multiplied together with the Gaussian likelihood term $p(\mathbf{y}|\mathbf{u})$, would result in $q(\mathbf{u})$ or $r(\mathbf{u})$ being an hybrid distribution, with some components taking values only in $\mathcal{A}$ and some other components taking real values. In such a case, evaluating the moments $\mathbb{E}[\boldsymbol{\phi}(\mathbf{u})]$ would be an issue. On the other hand, while many other statistics can be included in the vector $\boldsymbol{\phi}(\mathbf{u})$, e.g. cross moments of the form $u_i u_j$ for some or all pairs of variables, we will show in the experimental results session that our choice in (24) drives a robust and accurate MIMO detector. For instance, in the experimental section we show that the EC-based MIMO detector average mutual information in (7) is very close to the optimal detector for an scenario where the true posterior can be evaluated. Hence, there is little room for improvement of the EC solution by including higher order moments in $\boldsymbol{\phi}(\mathbf{u})$.

V. Optimizing the MIMO EC free energy

As described in the previous section, the goal in EC inference is to find $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ and $(\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s)$ such that $q(\mathbf{u})$ in (27), $r(\mathbf{u})$ in (31) (evaluated at $\boldsymbol{\gamma}_r = \boldsymbol{\gamma}_s - \boldsymbol{\gamma}_q$ and $\boldsymbol{\Lambda}_r = \boldsymbol{\Lambda}_s - \boldsymbol{\Lambda}_q$) and $s(\mathbf{u})$ in (35) satisfy

$$\mathbb{E}_q[u_i] = \mathbb{E}_r[u_i] = \mathbb{E}_s[u_i] \quad (36)$$

$$\mathbb{E}_q[u_i^2] = \mathbb{E}_r[u_i^2] = \mathbb{E}_s[u_i^2] \quad (37)$$

for $i \in [2m]$.

To achieve such a point, we present two algorithms. The so-called *single loop* (SL), iteratively updates either $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ or $(\boldsymbol{\gamma}_r, \boldsymbol{\Lambda}_r)$ and follows a message-passing procedure. The resulting algorithm has approximately the MMSE complexity per iteration (see Table I). On the other hand, by exploiting the fact that the EC free energy in (21) is a convex function w.r.t. $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$, the so-called *double loop* algorithm (DL) performs iteratively a convex optimization to set $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ for fixed $(\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s)$ to then update the latter. Simulation results in Section V-C show that the DL algorithm typically converges to a point closer to the stationarity conditions in (36)-(37). As a caveat, its complexity is extremely large (see Table I) and we would rather use it as a benchmark to improve the single loop approach.

It is important to remark that, for both algorithms, convergence to (36)-(37) is not guaranteed [39]. Actually, in most cases we observe that both algorithms get stuck in a $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$ point for which these parameters do not change anymore but at the same time the moment matching (MM) condition is not fully met. Our goal is to design robust algorithms to optimize the EC free energy such that they converge to stable $(\boldsymbol{\lambda}_q, \boldsymbol{\lambda}_r)$ points that are as close to the MM condition as possible.

A. The EC MIMO detector with single loop updates

We initialize $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ such that $q(\mathbf{u})$ in (27) coincides with the MMSE Gaussian approximation, i.e., $\boldsymbol{\gamma}_q^{(0)} = \mathbf{0}$ and $\Lambda_{qi}^{(0)} = E_s^{-1}$ $\forall i \in [2m]$ [13], [24]. The main steps are summarized Algorithm 1. The complexity per iteration is dominated by the computation of the covariance matrix of the $q(\mathbf{u})$ distribution in (27) at step 1) of the algorithm. This complexity is $\mathcal{O}(m^3)$, but independent on the constellation size M. After the matrix inversion, computing the mean of $q(\mathbf{u})$ requires $\mathcal{O}(m^2)$ operations. Computing the $r(\mathbf{u})$ mean and variance in (32) and (33) requires $\mathcal{O}(mM)$ operations.

Algorithm 1 The EC MIMO detector with SL updates

Fix a damping factor β. Set maximum number of iterations $I_{\text{EC-S}}$. Set $\ell = 0$.
Initialize $\boldsymbol{\gamma}_q^{(0)} = \mathbf{0}$ and $\Lambda_{qi}^{(0)} = E_s^{-1}$ $i \in [2m]$.
repeat
1) Given $\boldsymbol{\gamma}_q^{(\ell-1)}, \boldsymbol{\Lambda}_q^{(\ell-1)}$, compute $\mathbb{E}_q[u_i]$ and $\mathbb{E}_q[u_i^2]$, $i \in [2m]$.
2) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that $\mathbb{E}_s[u_i] = \mathbb{E}_q[u_i]$ and $\mathbb{E}_s[u_i^2] = \mathbb{E}_q[u_i^2]$, $i \in [2m]$.
3) Update $\boldsymbol{\gamma}_r^{(\ell)} = \boldsymbol{\gamma}_s^{(\ell)} - \boldsymbol{\gamma}_q^{(\ell)}$, $\boldsymbol{\Lambda}_r^{(\ell)} = \boldsymbol{\Lambda}_s^{(\ell)} - \boldsymbol{\Lambda}_q^{(\ell)}$.
4) Given $\boldsymbol{\gamma}_r^{(\ell)}, \boldsymbol{\Lambda}_r^{(\ell)}$, compute $\mathbb{E}_r[u_i]$ and $\mathbb{E}_r[u_i^2]$, $i \in [2m]$.
5) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that $\mathbb{E}_s[u_i] = \mathbb{E}_r[u_i]$ and $\mathbb{E}_s[u_i^2] = \mathbb{E}_r[u_i^2]$, $i \in [2m]$.
6) Update

$$\boldsymbol{\gamma}_q^{(\ell)} = \beta\left(\boldsymbol{\gamma}_s^{(\ell)} - \boldsymbol{\gamma}_r^{(\ell)}\right) + (1-\beta)\boldsymbol{\gamma}_q^{(\ell-1)}$$
$$\boldsymbol{\Lambda}_q^{(\ell)} = \beta\left(\boldsymbol{\Lambda}_s^{(\ell)} - \boldsymbol{\Lambda}_r^{(\ell)}\right) + (1-\beta)\boldsymbol{\Lambda}_q^{(\ell-1)}$$

7) $\ell = \ell + 1$
until convergence (or $\ell > I_{\text{EC-S}}$)

The complexity of the rest of steps does not depend on the constellation and thus the complexity is $\mathcal{O}(m)$. Therefore, if the algorithm is run for $I_{\text{EC-S}}$ iterations, the final complexity is $\mathcal{O}(m^3 I_{\text{EC-S}} + m^2 I_{\text{EC-S}} + mM I_{\text{EC-S}} + m I_{\text{EC-S}})$.

Numerical issues arise due to the fact that we are propagating moments between a continuous and a discrete distribution, particularly in scenarios where all the mass of the marginal $r(u_i)$ distribution is concentrated in a small region of a potentially very large QAM constellation. This leads to small values of the marginal variance $\text{Var}_r[u_i]$ and, consequently, Λ_{si} may diverge in step 5). In order to avoid numerical issues, we implement a *damping* (low-pass filter) in the update of $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ at step 6) of Algorithm 1. Smoothing parameter updates via damping is a fairly common technique to stabilize approximate inference iterative algorithms. See for instance [47]–[49] for discussions on message-passing stabilization.

B. The EC MIMO detector with double loop updates

The double loop algorithm is based on a simultaneous update of both $q(\mathbf{u})$ and $r(\mathbf{u})$ at every iteration by solving the following convex optimization problem for a fixed $(\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s)$

$$(\boldsymbol{\gamma}_q^*, \boldsymbol{\Lambda}_q^*) = \arg \min_{(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)} \log Z_{\text{EC}}(\boldsymbol{\gamma}_q, \boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_q, \boldsymbol{\Lambda}_s) \tag{38}$$
$$= \arg \min_{(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)} (\log Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q) + \log Z_r(\boldsymbol{\gamma}_s - \boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_s - \boldsymbol{\Lambda}_q))$$

At $(\boldsymbol{\gamma}_q^*, \boldsymbol{\Lambda}_q^*)$, both $q(\mathbf{u})$ and $r(\mathbf{u})$ have the same moments. Then, $(\boldsymbol{\gamma}_s, \boldsymbol{\Lambda}_s)$ is recomputed to enforce moment matching (as in step 2) of Algorithm 1). Instead of using the distribution $s(\mathbf{u})$ to iteratively communicate the moments between $q(\mathbf{u})$ and $r(\mathbf{u})$, as the single loop algorithm does, note that the double loop is directly optimizing together both $q(\mathbf{u})$ and $r(\mathbf{u})$ to then update $s(\mathbf{u})$. The main steps are outlined in Algorithm 2. We could use standard gradient descend to numerically solve (38) in step 1). Note that in (28), evaluating the gradient of $\log Z_q(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$ w.r.t. $(\boldsymbol{\gamma}_q, \boldsymbol{\Lambda}_q)$, requires a matrix inversion and a matrix product and thus a complexity of $\mathcal{O}(m^3 + m^2)$. If D denotes the number of gradient descend steps and $I_{\text{EC-D}}$ is the number of iterations, then the complexity is $\mathcal{O}(m^3 D I_{\text{EC-D}} + m^2 D I_{\text{EC-D}} + m I_{\text{EC-D}})$.

Algorithm 2 The EC MIMO detector with DL updates

Fix a damping factor β. Set maximum number of iterations $I_{\text{EC-D}}$. Set $\ell = 0$.
Initialize $\boldsymbol{\gamma}_s^{(0)} = \mathbf{0}$ and $\Lambda_{si}^{(0)} = E_s^{-1}$ $i \in [2m]$.
repeat
1) Given $\boldsymbol{\gamma}_s^{(\ell-1)}, \boldsymbol{\Lambda}_s^{(\ell-1)}$, solve the convex optimization in (38).
2) Compute $\boldsymbol{\gamma}_s^{(\ell)}, \boldsymbol{\Lambda}_s^{(\ell)}$ such that $\mathbb{E}_s[u_i] = \mathbb{E}_q[u_i]$ and $\mathbb{E}_s[u_i^2] = \mathbb{E}_q[u_i^2]$, $i \in [2m]$.
3) Update

$$\boldsymbol{\gamma}_s^{(\ell)} = \beta \left(\boldsymbol{\gamma}_s^{(\ell)}\right) + (1-\beta)\boldsymbol{\gamma}_s^{(\ell-1)}$$
$$\boldsymbol{\Lambda}_s^{(\ell)} = \beta \left(\boldsymbol{\Lambda}_s^{(\ell)}\right) + (1-\beta)\boldsymbol{\Lambda}_s^{(\ell-1)}$$

4) $\ell = \ell + 1$
until convergence (or $\ell > I_{\text{EC-D}}$)

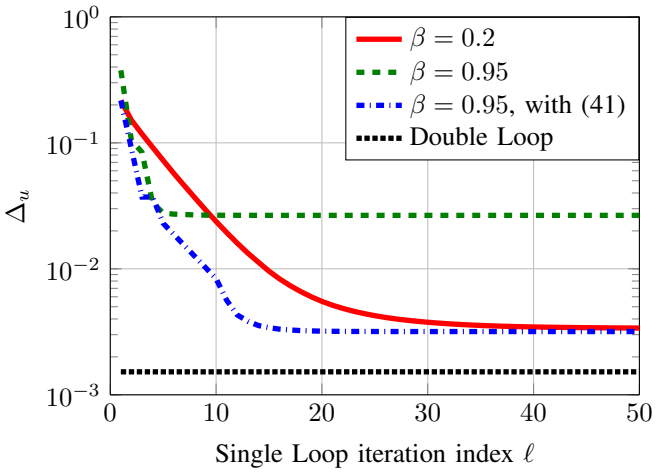
C. Assessing convergence

The moment matching condition in (36) and (37) represents the optimal operational point of the EC approximation. We emphasize that this notion of optimality is measured in terms of moment matching between tractable approximations to $p(\mathbf{u}|\mathbf{y})$ ($q(\mathbf{u})$ and $r(\mathbf{u})$ respectively), and not w.r.t. the distribution $p(\mathbf{u}|\mathbf{y})$ itself.

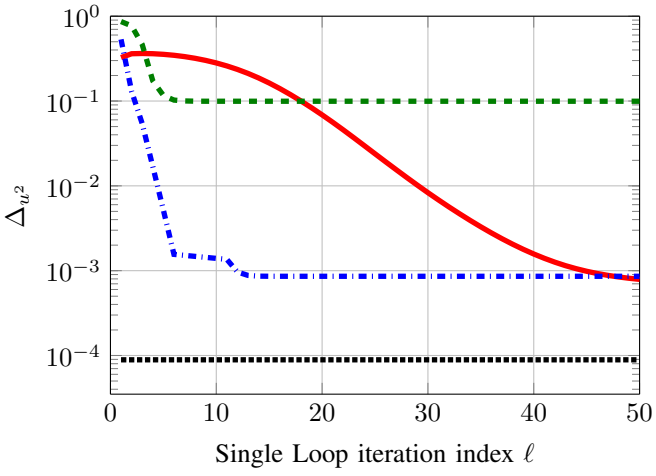
For our experiments, we study the evolution of the following two quantities along iterations of the single loop EC MIMO detector:

$$\Delta_u = \frac{1}{2m} \sum_{i=1}^{2m} \left| \mathbb{E}_q[u_i] - \mathbb{E}_r[u_i] \right|, \tag{39}$$

$$\Delta_{u^2} = \frac{1}{2m} \sum_{i=1}^{2m} \left| \mathbb{E}_q[u_i^2] - \mathbb{E}_r[u_i^2] \right|. \tag{40}$$



(a)



(b)

Fig. 3. We represent Δ_u and Δ_{u^2} for an 5×5 scenario with QPSK modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$.

In Fig. 3 we represent Δ_u and Δ_{u^2} for a 5×5 scenario with QPSK modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$. According to Fig. 2, this SNR value is far from the saturation

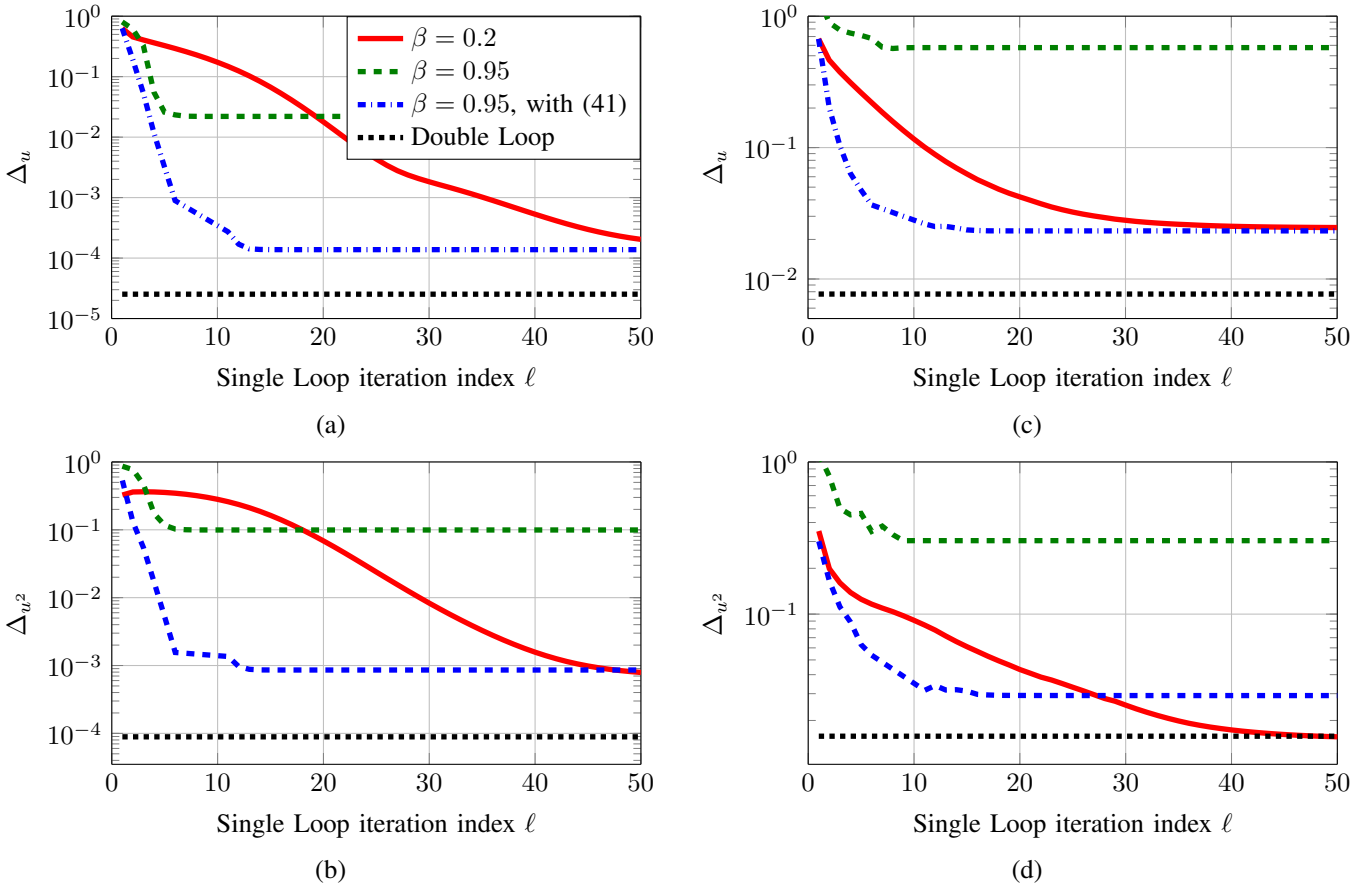


Fig. 4. In (a)-(b), we represent Δ_u and Δ_{u^2} for an 32×32 scenario with QPSK modulation at a SNR of 6dB, averaged over 10^4 realizations of both the channel matrix $\mathbf{H}$ and received vector $\mathbf{y}$. In (c)-(d), we reproduce the results for an 32×32 scenario with 64-QAM modulation at a SNR of 21dB.

regime (largest gap to channel capacity), and it is in this range where we aim the EC detector at substantially improving state-of-the-art methods. With dotted black line we represent the double loop benchmark, computed for $I_{\text{EC-D}} = 50$ iterations. At every iteration, we found that D, the number of gradient descend updates at step 1) of Algorithm 2, has to be to a very large value until the gradient norm was below a threshold of 0.1. We set an upper limit of $D = 2000$ and a gradient descend step-size of 10^{-3}. We remark that every gradient descend step is as complex as a single iteration of the single loop EC algorithm.

Three implementations of the SL algorithm are compared in Fig. 3. For the red solid line we have used $\beta = 0.2$, i.e., a very slow parameter update in step 6) of Algorithm 1. The opposite case is represented by the green dashed line, which has been computed with $\beta = 0.95$. While the $\beta = 0.2$ case approaches the double loop solution, achieving Δ_u and Δ_{u^2} around 10^{-3}, it requires in average 25 iterations to converge to such a stationary point. Recall that each single loop iteration is as complex as computing the MMSE estimate, due to the matrix inversion in (27). On the other hand, the $\beta = 0.95$ case quickly saturates (around 10 iterations), but its solution is still far from the MM condition.

In order to achieve a better trade-off between accuracy and complexity, we maintain the fast updates using $\beta = 0.95$, but modify the parameter update in Algorithm 1 and introduce a gradual decrease in the variance per component allowed at each iteration. More precisely, we set an iteration-dependent minimum value of the variance $\mathbb{E}_s[u_i^2]$ at step 5) of Algorithm 1 of the following form:

$$\text{Var}_s[u_i] = \max\left(2^{-\max(\ell-4,1)}, \text{Var}_r[u_i]\right), \qquad (41)$$

namely during the first 5 iterations we set a reasonably minimum high variance per component (0.5) and, from iteration 4, we let this minimum value to decrease exponentially fast with ℓ. The convergence of this implementation of the EC algorithm is represented in Fig. 3 with blue dashed-dotted lines. Observe that an improvement is achieved w.r.t. the $\beta = 0.95$ case, reducing the gap w.r.t. to the stationary point achieved by $\beta = 0.2$, without a significant penalty in speed of convergence, as it typically converges in less than 10 iterations. These effects are even more evident when we move to higher-dimensional scenarios. In Fig. 4 we consider a 32×32 scenario with QPSK (a)-(b) and 64-QAM modulation (c)-(d). Convergence speed is actually maintained and the gap w.r.t. the $\beta = 0.2$ case is clearly reduced. While the parameter update in (41) was obtained heuristically after an intense empirical evaluation of the algorithms, we interpret the improvement achieved as follows. Setting a high-variance parameter during the first iterations of the algorithm is crucial in the low-SNR regime in

TABLE I: Complexity order of different $r \times m$ MIMO detectors. In iterative algorithms, I_X denotes the number of iterations. D is the number of gradient descend steps for the double-loop EC detector.

MIMO detector	Complexity order
Optimal detector	M^m
MMSE	$m^3 + m^2 + mM$
soft MMSE-SIC [25]	$\mathcal{O}(m^3 + m^2 + mr^3 + mr^2 + mM)$
GTA [26]	$m^3 + m^2 M$
CHEMP [28]	$rm^2\ I_{\text{CHEMP}}$
EC (Single L.)	$m^3 I_{\text{EC-S}} + m^2 I_{\text{EC-S}} + mM I_{\text{EC-S}} + m I_{\text{EC-S}}$
EC (Double L.)	$m^3 D I_{\text{EC-D}} + m^2 D I_{\text{EC-D}} + m I_{\text{EC-D}}$

order to avoid over-fitting. For large values of β, we observed that the single loop EC algorithm performance is degraded by very small values of the $r(u_i)$ variance ($\text{Var}_r[u_i]$) at early iterations (step 4) of Algorithm 1, indicating a very peaky distribution around a small region of the QAM constellation. Note that a very small variance is propagated to the $s(u_i)$ distribution at step 5) of Algorithm 1 with very large values of Λ_{si}. According to (35), we have

$$\Lambda_{si}^{-1} = \text{Var}_s[u_i] = \mathbb{E}_r[u_i^2] - (\mathbb{E}_r[u_i])^2 = \text{Var}_r[u_i], \qquad (42)$$

and the same effect is propagated to Λ_{qi} at step 6) of the algorithm unless β is small enough. Very large values of Λ_{qi} will dominate the diagonal of the matrix in (27) and, ultimately, this implies that successive steps of the EC algorithm will not be able to significantly change the u_i marginal distribution anymore. Note that this is dramatic to the algorithm performance if the mode of the $r(u_i)$ distribution is placed at the wrong symbol, which is likely to happen at high-noise levels.

Instead of using small values of β to control sudden changes in parameter updates, with the update in (41), we propose an easy way to artificially control overconfident distributions at early steps of the algorithm, which would restrain the EC algorithm to move far away from the MMSE initial estimate. We note that using the EC moment matching criterion many other variants of the single loop update methods can be tested and compared with our proposal. However, no significant differences have been appreciated when we measure the system performance in terms of the mutual information in (7) or system bit error rate (BER). In the rest of the paper, regardless of the dimension of the system or constellation order, we implement the EC detector using the single loop approach with $\beta = 0.95$, the progressive variance limit in (41) and a maximum number of iterations of $I_{\text{EC-S}} = 10$.

D. Complexity

In Table V-D we summarize the main complexity order of the algorithms presented and those that will be used in our simulation experiments in the next section. In iterative algorithms, I_X denotes the number of iterations. As a rule of thumb, if we run the EC MIMO detector using $I_{\text{EC-S}} = 10$ iterations, the incurred complexity is around 10 times larger than the MMSE, GTA and CHEMP complexities. However, the significant gain in performance that we report in the next section can justify the increased-complexity of the proposed EC detector.

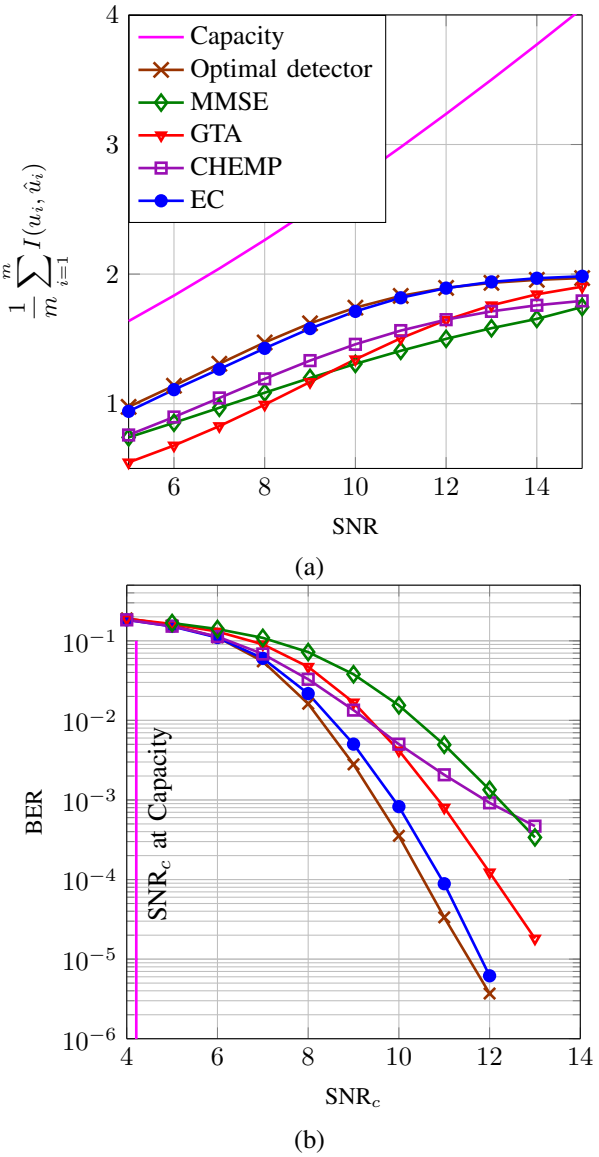


Fig. 5. For an 5×5 MIMO system with QPSK modulation, in (a) we show the achievable transmission rates. In (b), we include simulated performance when a $(3, 6)$-regular LDPC code with block length 5120 bits is used.

VI. Experimental Results

In the following, we include simulation performance results that demonstrate the accuracy of the EC approximation. In our experiments, we compare our proposal with the soft output MMSE solution in [13], [24], the soft version of the MMSE-SIC in [25], the GTA algorithm in [26], and the CHEMP method in [28]. To avoid cluttering, we do not include in our experiments the GMPID algorithm [30], since it performs close to CHEMP. For similar reasons, we do not include the EP method proposed in [33], since it performs similarly to GTA when used for probabilistic detection [36].

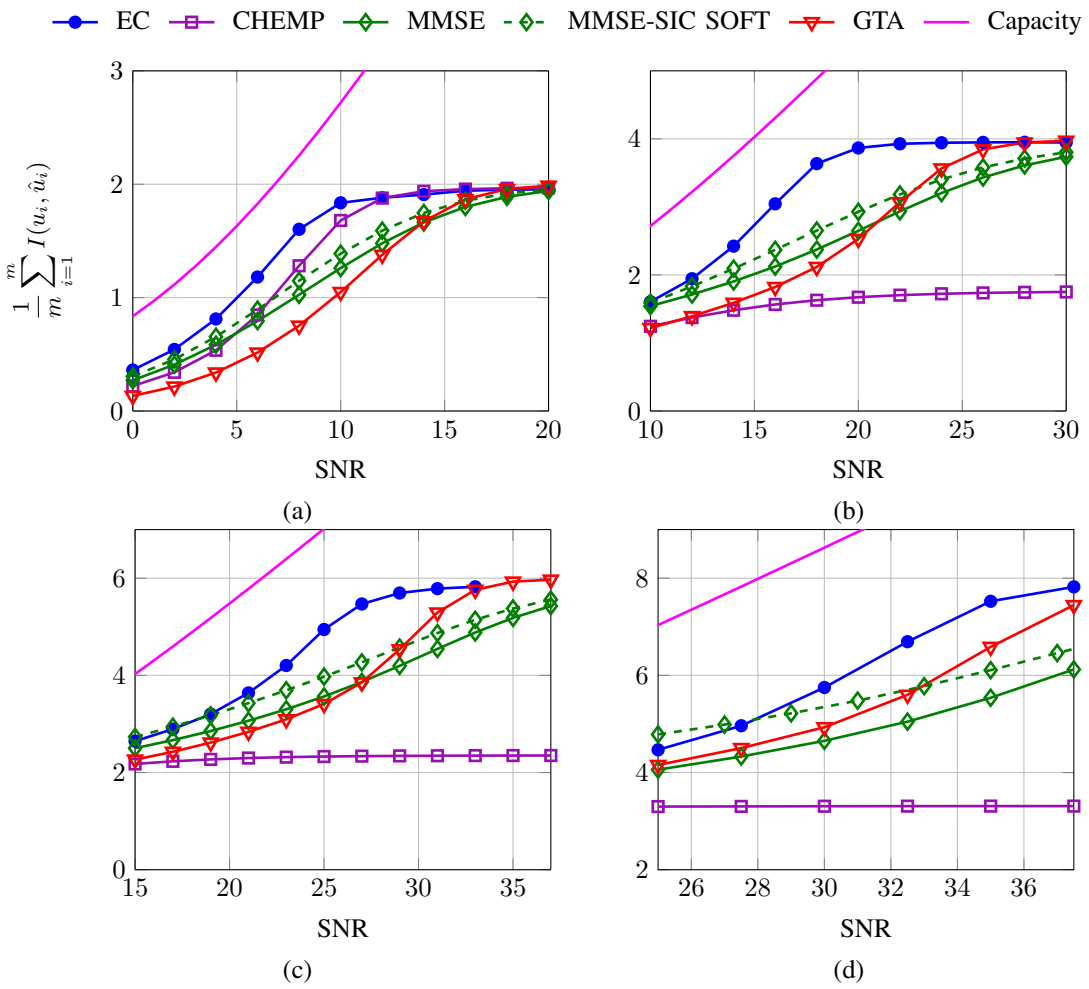


Fig. 6. Transmission rate computed for an $m = r = 32$ MIMO system and different constellation orders with QPSK modulation (a), 16-QAM modulation (b), 64-QAM modulation (c) and 256-QAM modulation (d).

A. A Low Dimensional MIMO System

Consider again the 5×5 scenario with QPSK modulation described in Fig. 2. Recall that the dimensionality is small enough so we are able to solve the marginalization in (8) exactly, which represents the optimal detector. In Fig. 5(a) we include now the results for the EC MIMO detector. Remarkably, it essentially overlaps the optimal detector performance, achieving a large gain w.r.t. GTA, MMSE and CHEMP. When the number of antennas is small (5 in our case), the columns of the channel matrix $\mathbf{H}$ are typically non-orthogonal and this limits the MMSE performance [13], [24]. Also, the CHEMP method relies on the matrix $m^{-1}\mathbf{H}^{\top}\mathbf{H}$ being diagonal and for a small m, this assumption is unrealistic [28].

Results in Fig. 5(a) indicate that the MIMO system performance will highly benefit from the more accurate estimates to the symbol posterior marginals $p(u_i|\mathbf{y})$ provided by the EC detector. To corroborate this fact, we augment the system model in Fig. 1 by including an LDPC channel encoding stage at the transmitter and an LDPC channel decoder at the receiver. The LDPC channel decoder is fed by soft coded bit probabilities computed using the symbol posterior marginals $p(u_i|\mathbf{y})$ (or their estimates), according to the bit-modulation mapping. It is well known that the more accurate the probabilistic detector is, the better performance is obtained after the LDPC decoding stage using BP [24], [50], [51]. In Fig. 5(b), we show for this scenario the simulated BER measured after the LDPC decoding stage (solid lines). A $(3, 6)$-regular LDPC code with block length equal to 5120 bits has been used. Note that, to simulate the coded performance, the SNR definition in (2) is corrected by the coding rate R (the coding rate is $R = 0.5$ in the case of $(3, 6)$-regular LDPC code). To avoid confusion, we denote this by SNR_c, and thus SNR_c(dB)=SNR+$10\log_{10}(R)$. Results have been averaged over 5000 realizations of the channel matrix $\mathbf{H}$. In terms of coded performance, the gap between optimal detection and EC is only about 0.4 dB measured at a BER of 10^{-4} while the gap to GTA is over 1.5 dB. In all scenarios observe that, while the soft MMSE-SIC method always improves MMSE, and also GTA al low SNR values, its performance is still far from the EC detector.

B. A 32×32 MIMO system

In a larger scenario, exact marginalization is not viable anymore and we fully rely on approximate methods. In Fig. 6, we represent the obtained achievable rates for a 32×32 MIMO system using QPSK modulation (a), 16-QAM modulation (b), 64-QAM modulation (c), and 256-QAM modulation (d).

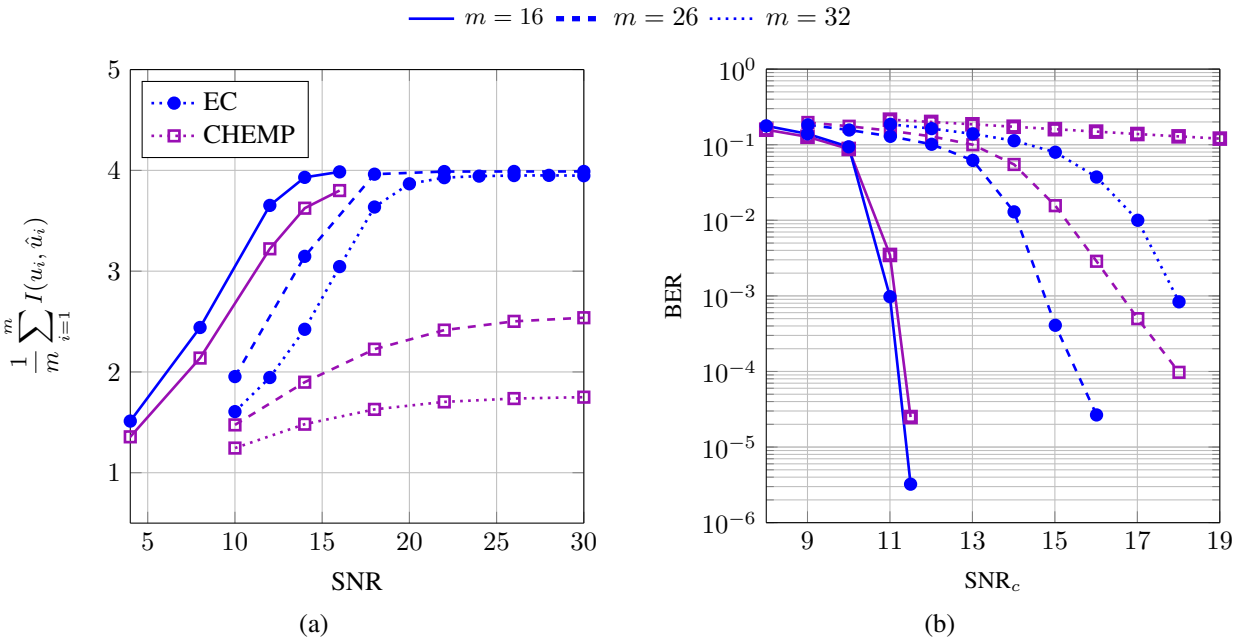


Fig. 7. For an $32 \times m$ MIMO system with 16-QAM modulation, in (a) we show the achievable transmission rates for different m values. In (b), we include simulated performance when a $(3, 6)$-regular LDPC code with block length 5120 bits is used.

While CHEMP and EC are competitive for the QPSK case, CHEMP is no longer a viable option in the 16-QAM or 64-QAM cases. As discussed in [28], the variance of the interference noise that CHEMP aims to iteratively cancel grows with the constellation order. For $m = r$ and high order constellations the interference noise becomes excessively large. Note that the soft MMSE-SIC method always improves MMSE and GTA al low-intermediate SNR values but still its performance is far from the EC detector.

Following [28], it can be checked that CHEMP becomes effective again as we reduce the number of transmitting antennas, i.e., if $m < r$. In Fig. 7 (a), we compare the EC and CHEMP transmission rates for a 16-QAM modulation with $r = 32$ and $m = 16$, 26 and 32. In (b), we include BER simulation results using the $(3, 6)$-regular LDPC code with block length equal to 5120 bits. For small m values, CHEMP is comparative to the EC solution. However, its performance is severely degraded as m approaches r. CHEMP can be regarded as a Gaussian message-passing distributed implementation of the EC algorithm for those cases where interference is "locally" tractable. Unlike CHEMP, the EC algorithm performs the update of all parameters at the same time in a centralized manner. These results show that EC MIMO detector is robust against the increase in the constellation order. In the following we solely consider $m = r$ scenarios with high order constellations and hence we omit CHEMP from the results.

We complete the study of this scenario by including BER performance results using LDPC constructions that are designed to improve the performance of the $(3, 6)$-regular LDPC code used in previous experiments. In Fig. 8 with dashed lines we show the performance of the rate-$1/2$ irregular LDPC code in [6, Example 3.99] with block length equal to 30720 bits. We also include simulation results (solid lines) for a convolutional LDPC (LDPCC) code constructed by spatially-coupling 48 independent copies of a $(3, 6)$-regular LDPC code, each having block length of 640 bits, with low-rate terminations [52]. The resulting coding rate is 0.479 and the total block length is 30720 bits. For the irregular LDPC code, at moderate SNR EC is able to provide a significant gain, which vanishes at high SNR because of the LDPC error floor. In contrast, because the LDPCC code has large minimum distance, no error floor has been observed in the range of SNR considered and EP achieves a stable gain of 2.5 dB with respect to GTA. Finally, with dotted lines we include simulation results for a LDPCC code[3] with the same block length but constructed by spatially-coupling 48 independent copies of a $(3, 24)$-regular LDPC code. The resulting coding rate is 0.869.

VII. Conclusions

Probabilistic symbol detection is a fundamental problem in high-dimensional MIMO communications since the accuracy of the method employed to approach the true posterior solution may bring significant performance gains when combined with a modern capacity-approaching channel coding scheme. In this paper we have shown how the EC approximate inference methodology, when applied to the posterior probability distribution of the transmitted symbols, can lead to accurate estimates of the marginal distribution for each transmitted symbol. Further, by computing the average per-antenna mutual information between the transmitted symbols and those distributed according to the EC output, we have shown that the system achievable rate heavily depends on the probabilistic detector accuracy and thus the importance of this stage cannot

[3]LDPCC codes are generated using protographs [53] in order to optimize its minimum distance, as described in [54].

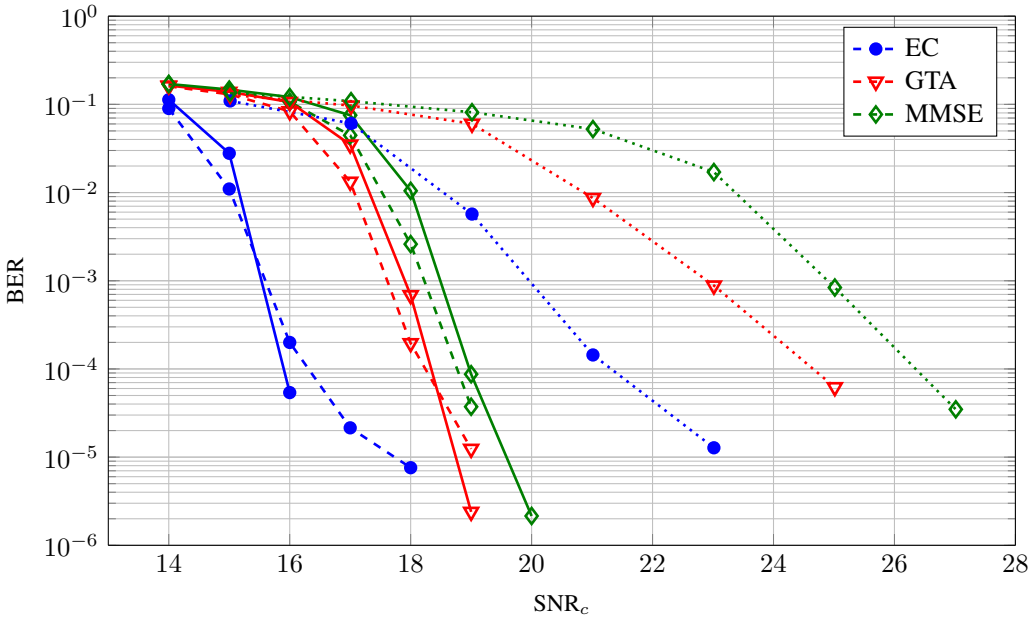


Fig. 8. System performance of an 32×32 16-QAM using the irregular rate-$1/2$ LDPC code in [6, Example 3.99] with (dashed lines) block length 30720 bits, a $(3, 6)$-regular LDPC convolutional code (solid lines) with the same block-length and coding rate 0.479, and a $(3, 24)$-regular LDPC convolutional code (dotted lines) with the same block-length and coding rate 0.8698 [54].

be diminished by using a more powerful channel code. This is actually corroborated by testing the system performance when we combine the probabilistic output of the symbol detectors with an LDPC channel decoder based on belief propagation. The presented EC probabilistic MIMO detector has cubic complexity with the number of antennas and it is able to greatly improve state-of-the-art methods within only 10 iterations, where a matrix inversion has to be performed per iteration.

REFERENCES

[1] J. Mietzner, R. Schober, L. Lampe, W. H. Gerstacker, and P. A. Hoeher, "Multiple-antenna techniques for wireless communications - a comprehensive literature survey," *IEEE Communications Surveys Tutorials*, vol. 11, pp. 87–105, June 2009.

[2] L. Zheng, P. Viswanath, and D. N. C. Tse, "Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels," *IEEE Transactions on Information Theory*, vol. 49, pp. 1073–1095, May 2003.

[3] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo-codes," in *Proc. IEEE International Conference on Communications, Geneva, Switzerland*, May 1993.

[4] T. J. Richardson and R. Urbanke, *Modern coding theory*. Cambridge University Press, 2008.

[5] F. R. Kschischang, B. J. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, pp. 498–519, Feb. 2001.

[6] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1566–1577, June 2005.

[7] Z. Guo and P. Nilsson, "Algorithm and implementation of the K-best sphere decoding for MIMO detection," *IEEE Journal on Selected Areas in Communications*, vol. 24, pp. 491–503, March 2006.

[8] G. D. Golden, C. J. Foschini, R. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electronics Letters*, vol. 35, pp. 14–16, January 1999.

[9] T.-h. Liu and Y.-L. Liu, "Modified fast recursive algorithm for efficient MMSE-SIC detection of the V-BLAST system," *IEEE Transactions on Wireless Communications*, vol. 7, pp. 3713–3717, October 2008.

[10] H. Zhao, H. Long, and W. Wang, "Tabu Search Detection for MIMO Systems," in *Proc. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, Athens, Greece*, September 2007.

[11] N. Srinidhi, T. Datta, A. Chockalingam, and B. S. Rajan, "Layered Tabu Search Algorithm for Large- MIMO Detection and a Lower Bound on ML Performance," *IEEE Transactions on Communications*, vol. 59, pp. 2955–2963, November 2011.

[12] Q. Zhou and X. Ma, "Element-Based Lattice Reduction Algorithms for Large MIMO Detection," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 274–286, 2013.

[13] G. Caire, R. R. Muller, and T. Tanaka, "Iterative multiuser joint decoding: optimal power allocation and low-complexity implementation," *IEEE Transactions on Information Theory*, vol. 50, pp. 1950–1973, September 2004.

[14] J. Goldberger, "Improved MIMO Detection based on Successive Tree Approximations," in *Proc. 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey*, June 2013.

[15] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMO," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 1941–1988, April 2015.

[16] J. Boutros, N. Gresset, L. Brunel, and M. Fossorier, "Soft-input soft-output lattice sphere decoder for linear channels," in *Proc. IEEE Global Communications Conference, San Francisco, USA*, December 2003.

[17] C. Studer, A. Burg, and H. Bolcskei, "Soft-output sphere decoding: algorithms and VLSI implementation," *IEEE Journal on Selected Areas in Communications*, vol. 26, pp. 290–300, February 2008.

[18] R. Wang and G. B. Giannakis, "Approaching MIMO channel capacity with soft detection based on hard sphere decoding," *IEEE Transactions on Communications*, vol. 54, pp. 587–590, April 2006.

[19] B. Steingrimsson, Z.-Q. Luo, and K. M. Wong, "Soft quasi-maximum-likelihood detection for multiple-antenna wireless channels," *IEEE Transactions on Signal Processing*, vol. 51, pp. 2710–2719, November 2003.

[20] T. Datta, N. A. Kumar, A. Chockalingam, and B. S. Rajan, "A Novel Monte-Carlo-Sampling-Based Receiver for Large-Scale Uplink Multiuser MIMO Systems," *IEEE Transactions on Vehicular Technology*, vol. 62, pp. 3019–3038, September 2013.

[21] M. Hansen, B. Hassibi, A. G. Dimakis, and W. Xu, "Near-Optimal

Detection in MIMO Systems Using Gibbs Sampling," in *Proc. IEEE Global Telecommunications Conference, Hawaii, USA*, November 2009.

[22] R.-R. Chen, R. Peng, A. Ashikhmin, and B. Farhang-Boroujeny, "Approaching MIMO capacity using bitwise Markov Chain Monte Carlo detection," *IEEE Transactions on Communications*, vol. 58, pp. 423–428, February 2010.

[23] Y. Jia, C. Andrieu, R. J. Piechocki, and M. Sandell, "Improving soft output quality of MIMO demodulation algorithm via importance sampling," in *Proc. IEE International Conference on 3G Mobile Communication Technologies, London, UK*, 2004.

[24] A. Sanderovich, M. Peleg, and S. Shamai, "LDPC coded MIMO multiple access with iterative joint decoding," *IEEE Transactions on Information Theory*, vol. 51, pp. 1437–1450, April 2005.

[25] J. Wang and S. Li, "Soft versus hard interference cancellation in MMSE OSIC MIMO detector: A comparative study," in *Proc. 2007 4th International Symposium on Wireless Communication Systems, Trondheim, Norway*, October 2007.

[26] J. Goldberger and A. Leshem, " MIMO Detection for High-Order QAM Based on a Gaussian Tree Approximation," *IEEE Transactions on Information Theory*, vol. 57, pp. 4973–4982, August 2011.

[27] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-Theoretically Optimal Compressed Sensing via Spatial Coupling and Approximate Message Passing," *IEEE Transactions on Information Theory,*, vol. 59, pp. 7434–7464, November 2013.

[28] T. L. Narasimhan and A. Chockalingam, "Channel Hardening-Exploiting Message Passing (CHEMP) Receiver in Large-Scale MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 847–860, October 2014.

[29] L. Liu, C. Yuen, Y. L. Guan, and Y. Li, "Capacity-achieving iterative LMMSE detection for MIMO-NOMA systems," in *2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia*, May 2016.

[30] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and Y. Su, "Convergence analysis and assurance for gaussian message passing iterative detector in massive MU-MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 6487–6501, September 2016.

[31] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. 2015 IEEE International Symposium on Information Theory, Hong Kong, China*, June 2015.

[32] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian Message Passing Iterative Detection for MIMO-NOMA Systems with Massive Access," in *2016 IEEE Global Communications Conference (GLOBECOM), Washington DC, USA*, Dec 2016.

[33] J. Céspedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Expectation Propagation Detection for High-Order High-Dimensional MIMO Systems," *IEEE Transactions on Communications*, vol. 62, pp. 2840–2849, August 2014.

[34] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. of the Seventeenth Coference on Uncertainty in Artificial Intelligence, Seattle, USA*, August 2001.

[35] M. W. Seeger, "Expectation propagation for exponential families," tech. rep., 2005.

[36] J. Cespedes, P. M. Olmos, M. Sánchez-Fernández, and F. Perez-Cruz, "Improved performance of LDPC-coded MIMO systems with EP-based soft-decisions," in *Proc. 2014 IEEE International Symposium on Information Theory, Hawaii, USA*, June 2014.

[37] I. Santos, J. J. Murillo-Fuentes, R. Boloix-Tortosa, E. A. de Reyna, and P. M. Olmos, "Expectation Propagation as Turbo Equalizer in ISI Channels," *IEEE Transactions on Communications*, vol. 65, pp. 360–370, January 2017.

[38] G. M. Vitetta, D. P. Taylor, G. Colavolpe, F. Pancaldi, and P. A. Martin, *Wireless Communications: Algorithmic Techniques*. John Wiley & Sons, Ltd, 2013.

[39] M. Opper and O. Winther, "Expectation Consistent Approximate Inference," *Journal of Machine Learning Research*, vol. 6, pp. 2177–2204, December 2005.

[40] T. J. Richardson, M. A. Shokrollahi, and R. Urbanke, "Design of capacity approaching irregular low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 47, pp. 619–637, February 2001.

[41] D. J. Costello, Jr., L. Dolecek, T. Fuja, J. Kliewer, D. G. M. Mitchell, and R. Smarandache, "Spatially coupled sparse codes on graphs: theory and practice," *IEEE Communications Magazine*, vol. 52, pp. 168–176, July 2014.

[42] S. Kudekar, T. Richardson, and R. Urbanke, "Spatially Coupled Ensembles Universally Achieve Capacity under Belief Propagation," *IEEE Transactions on Information Theory*, vol. 59, pp. 7761–7813, December 2013.

[43] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends in Machine Learning, 2008.

[44] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunication*, vol. 10, pp. 585–596, November 1999.

[45] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag, New York, 2006.

[46] J. Ketonen, M. Juntti, and J. R. Cavallaro, "Performance-complexity comparison of receivers for a LTE MIMO OFDM system," *IEEE Transactions on Signal Processing*, vol. 58, pp. 3360–3372, June 2010.

[47] T. Heskes, "Stable fixed points of loopy belief propagation are minima of the Bethe free energy," in *Proc. 2002 Advances in Neural Information Processing Systems*, vol. 14, MIT Press, 2003.

[48] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 53, pp. 4422–4437, December 2007.

[49] G. Elidan, I. McGraw, and D. Koller, "Residual belief propagation: informed scheduling for asynchronous message passing," in *Proceedings of the Twenty-second Conference on Uncertainty in Artificial Intelligence, Cambridge, USA*, July 2006.

[50] P. M. Olmos, J. J. Murillo-Fuentes, and F. Pérez-Cruz, "Joint nonlinear channel equalization and soft LDPC decoding with Gaussian processes," *IEEE Transactions on Signal Processing*, vol. 58, pp. 1183–1192, March 2010.

[51] A. G. D. Uchoa, R. C. D. Lamare, and C. Healy, "Iterative Detection and Decoding Algorithms For Block-Fading Channels Using LDPC Codes," in *Proc. 2014 IEEE Wireless Communications and Networking Conference, Istanbul, Turkey*, April 2014.

[52] D. G. M. Mitchell, A. E. Pusane, M. Lentmaier, and D. J. Costello, Jr., "Exact Free Distance and Trapping Set Growth Rates for LDPC Convolutional Codes," in *Proc. IEEE International Symposium on Information Theory, St. Petersburg, Russia*, 2011.

[53] J. Thorpe, "Low-Density Parity-Check (LDPC) codes constructed from protographs," INP Progress Report 42-154, Jet Propulsion Laboratory, Pasadena, CA, 2003.

[54] D. Mitchell, M. Lentmaier, and D. J. Costello, Jr., "Spatially Coupled LDPC Codes Constructed From Protographs," *IEEE Transactions on Information Theory*, vol. 61, pp. 4866–4889, September 2015.